A STRATIFIED TWO-STAGE SAMPLING PLAN FOR THE ESTIMATION OF DISEASE INCIDENCE IN THE DAIRY COWS OF NEW YORK STATE

BU-104-M D. S. Robson February 9, 1959

Introduction

Two-stage sampling, or cluster-sampling with subsampling within clusters, is commonly employed in sample surveys to reduce both survey costs and sampling error. The population is conceived of as being partitioned into clusters as, for example, the human population is partitioned into family groups represented by households, and a sample is drawn in two stages; first, a random sample of clusters (e.g., households) is selected and then a random sample of individuals is selected from each of the chosen clusters. If individuals within clusters are more nearly alike than individuals from different clusters then this two-stage sampling plan is statistically more efficient than the simple one-stage plan.

Natural clusters occur in almost every biological population, and among domestic plants and animals the clustering is induced by the households of the human population. The population of dairy cows, in particular, partitions naturally into the individual dairy farm herds; animals within a herd are subjected to common management practices and may be expected to share common diseases through contact exposure. Animals within a herd should therefore be more alike in their disease history than animals from different herds, and the two-stage sample of herds and cows within herds should represent a near optimum sampling plan. Geographic stratification by county with proportionate allocation will further improve the sample if disease incidence varies from county to county.

The Sampling Plan

A sampling plan for a first survey into a new area of investigation should incorporate features which guarantee the estimability of sampling error. The maximum sampling error of an incidence estimate may be computed in advance on the basis of simple, one-stage binomial sampling theory, and may serve as a guide in determining the sample size necessary to yield the desired degree of precision in the incidence estimate. This approximation, however, leads to over sampling of the population if the gains in precision due to two-stage sampling

are substantial, or, equivalently, leads to underestimating the precision of the incidence estimate. The sampling plan proposed here for the initial incidence survey therefore specifies that the subsample from each selected herd shall include two cows, rather than one, to provide a measure of the within-herd variability. Total sample size for the initial survey is fixed somewhat arbitrarily at 1000; this sample is larger than would be required for most purposes, giving incidence estimates accurate to within 3 or 4% with high probability, but necessary for an accurate picture of the variability which contributes to sampling error. The sample of 500 herds is allocated to the counties in proportion to the relative frequency of dairy farms listed in the latest census of agriculture as shown in Table 1.

Sample size is also determined in part by the cost and the facilities available for the analysis of the sample. If the determination of disease is to be made by serological tests for antibodies then laboratory facilities will be a limiting factor with respect to sample size and laboratory costs will represent the major component of the total cost of analysis. In this case, consideration should also be given to the cost of storing the serum samples for future tests as new serological techniques are developed.

An Unbiased Estimation Procedure

The proportion of diseased animals in a sample consisting of n animals from each of k herds is an estimate of the average incidence per herd in the population; that is, if there are K herds in a given county with proportions of diseased animals equal to $p_1, p_2, \cdots, p_K$, respectively, then the incidence in the sample of kn animals from that county is an estimate of $\bar{p} = \frac{1}{K} \sum_{1}^{K} p_i$. This estimate, say $\hat{\bar{p}}$

$$\hat{\bar{p}} = \frac{1}{k} \sum_{1}^{k} \hat{p}_i$$

where $\hat{p}_i$ is the proportion of diseased animals in the sample of size n from the i'th herd, is subject to a sampling error of

$$\mathrm{Var}(\hat{\bar{p}}) = \frac{1}{Kk} \left[\sum_{1}^{K} \frac{N_i - n}{n(N_i - 1)} p_i(1-p_i) + \frac{K-k}{K-1} \sum_{1}^{K} (p_i - \bar{p})^2 \right]$$

where N_i is the total number of animals in the i'th herd. Sampling variance is made up of two components, the first due to sampling within a herd and the second due to the sampling of herds. These two components, say V_1 and V_2, may be estimated separately since

$$\hat{V}_1 = \frac{1}{k^2(n-1)} \sum_1^k \frac{N_i-n}{N_i-1} \hat{p}_i(1-\hat{p}_i)$$

estimates V_1 and

$$\hat{V}_2 = \frac{K-k}{Kk}\left[\frac{1}{k-1}\sum_1^k (\hat{p}_i-\hat{\bar{p}})^2 - k\hat{V}_1\right]$$

estimates V_2, so that $\mathrm{Var}(\hat{\bar{p}})$ is estimated by $\hat{V}_1+\hat{V}_2$.

The parameter $\bar{p}$, representing the average incidence per herd, is not the incidence for the total population unless the incidence p_i in each herd is the same or the size N_i of each herd is the same. Incidence p in the total population is, rather, the weighted average of the p_i,

$$p = \frac{1}{N}\sum_1^K N_i p_i$$

where $N = \sum_1^K N_i$ is the total number of cows in the county. The adjustment of the estimate $\hat{\bar{p}}$ required to produce an estimate of this parameter p is shown by

$$\hat{p} = \hat{\bar{p}} + \frac{K-1}{N(k-1)}\left[\sum_1^k N_i\hat{p}_i - \hat{\bar{p}}\sum_1^k N_i\right] .$$

Sampling error variance of this adjusted estimator takes a more complicated form, as may be seen by writing the estimator in the form

$$\hat{p} = p^* + \frac{1}{k}\sum_1^k(\hat{p}_i-p_i) + \frac{K-1}{N(k-1)}\left[\sum_1^k N_i(\hat{p}_i-p_i) - \frac{1}{k}\sum_1^k N_i \sum_1^k(\hat{p}_i-p_i)\right]$$

where

$$p^* = \frac{1}{k}\sum_1^k p_i + \frac{K-1}{N(k-1)}\left[\sum_1^k N_i p_i - \frac{1}{k}\sum_1^k N_i \sum_1^k p_i\right] .$$

In this form, $\hat{p}$ is expressed in terms of a statistic p^*, which is not actually observable due to incomplete sampling within herds, and a remainder representing the error component due to sampling within herds. The statistic p^* is the Hartley-Ross [1] ratio estimator of a proportion, the varaince of which has been computed by Robson [2]. Since p^* and the remainder are uncorrelated, the varaince of $\hat{p}$ is expressible as the variance of p^* plus a within herd variance component. While the exact formula for $\mathrm{Var}(\hat{p})$ and its unbiased estimator $\widehat{\mathrm{Var}}(\hat{p})$ are not yet available, they may be obtained directly through application of the methods given by Robson (see [2] or [3]).

A much simpler, alternative estimate of p is given by

$$p' = \frac{K}{Nk} \sum_{1}^{k} N_i \hat{p}_i$$

$$= \frac{K}{Nk} \sum_{1}^{k} N_i p_i + \frac{K}{Nk} \sum_{1}^{k} N_i (\hat{p}_i - p_i) \quad .$$

The sampling error variance of p' contains two components very similar in structure to the components V_1 and V_2 of $\mathrm{Var}(\hat{p})$; thus,

$$\mathrm{Var}(p') = \frac{K}{kN^2} \left\{ \sum_{1}^{K} \frac{N_i^2 (N_i - n)}{n(N_i - 1)} p_i (1 - p_i) + \frac{K-k}{K-1} \sum_{1}^{K} N_i^2 (p_i - p)^2 \right\} \quad .$$

These three estimators $\hat{\bar{p}}$, $\hat{p}$ and p' are related by the equation

$$\hat{p} = \hat{\bar{p}} \left[1 - \frac{K-1}{N(k-1)} \sum_{1}^{k} N_i \right] + \frac{k(K-1)}{K(k-1)} p' \quad .$$

An answer to the question as to which of the two statistics $\hat{p}$ and p' is the better estimator of p must await an evaluation of $\mathrm{Var}(\hat{p})$; however, the usual advantages of ratio-type estimation may be expected to prevail with the result that $\hat{p}$ is better than p'.

References

1 H. O. Hartley and A. Ross, "Unbiased ratio estimators," Nature, Vol. 174 (1954), p. 270.

2 D. S. Robson, "Some applications of multivariate polykays to the theory of unbiased ratio-type estimation," Journal of the American Statistical Association, Vol. 52 (1957), pp. 511-522.

3 D. S. Robson, "An unbiased sampling and estimation procedure for creel censuses of fishermen," No. BU-102-M of the Biometrics Unit mimeograph series, Cornell University (1959).

Table 1. Allocation of 1000 milk cow serum samples to 54 counties

		1954 Population		Proposed Sample	
		Farms	Cows	Farms	Cows
1	Albany	695	11,109	5	10
2	Allegany	1756	24,920	12	24
3	Broome	1440	23,567	10	20
4	Cattaraugus	2447	45,153	17	34
5	Cayuga	1449	23,060	10	20
6	Chautauqua	2915	42,141	20	40
7	Chemung	732	8,656	5	10
8	Chenango	1878	42,935	13	26
9	Clinton	1588	30,995	11	22
10	Columbia	904	20,008	6	12
11	Cortland	955	27,044	7	14
12	Delaware	2469	64,204	17	34
13	Dutchess	824	23,531	6	12
14	Erie	2169	30,314	15	30
15	Essex	641	7,075	4	8
16	Franklin	1415	27,159	10	20
17	Fulton	521	7,367	4	8
18	Genesee	1150	19,121	8	16
19	Greene	777	12,487	5	10
20	Herkimer	1445	36,290	10	20
21	Jefferson	2547	57,646	18	36
22	Lewis	1372	33,807	10	20
23	Livingston	1089	21,992	7	14
24	Madison	1532	33,029	11	22
25	Monroe	1172	17,836	8	16
26	Montgomery	1077	24,608	8	16
27	Niagara	1522	12,088	11	22
28	Oneida	2618	55,649	18	36
29	Onondaga	1485	27,698	10	20
30	Ontario	1401	19,149	10	20
31	Orange	1281	38,349	9	18
32	Orleans	827	9,109	6	12
33	Oswego	1963	24,056	14	28
34	Otsego	2375	49,732	17	34
35	Putnam	124	3,209	1	2
36	Rensselaer	1114	17,459	8	16
37	St.Lawrence	3816	77,666	27	54
38	Saratoga	1178	13,735	8	16
39	Schenectady	284	3,905	2	4
40	Schoharie	1208	24,725	8	16
41	Schuyler	579	5,295	4	8
42	Seneca	578	6,934	4	8
43	Steuben	2397	32,565	17	34
44	Suffolk	313	2,585	2	4
45	Sullivan	967	13,395	7	14
46	Tioga	1130	18,210	8	16
47	Tompkins	927	14,008	6	12
48	Ulster	864	11,648	6	12
49	Warren	337	1,474	2	4
50	Washington	1717	33,648	12	24
51	Wayne	1364	14,406	10	20
52	Wyoming	1567	33,154	11	22
53	Yates	687	8,054	5	10
	Total	71582	1,292,887	500	1000